מסמך מסכם – מיני פרויקט בנושאים במדעי הרוח הדיגיטליים

יעל חדד ומרק אילישייב

מטרת הפרויקט:

יצירת קישור וחיבור בין תיקי הארכיון למקורות מידע חיצוניים כמו imdb והצלבת המטה דאטה בין המקורות השונים.

מצב הסריקות בארכיון:

עבור העשורים 80-60 הסריקות רובן התבצעו, אך העשורים המאוחרים יותר – קיימים מלכתחילה תיקים מועטים באופן יחסי בארכיון הדיגיטלי , אך לא נתקלנו במקרים רבים בהם יש צורך להזמין סריקת תיק שעולה בחיפוש אל טרם נסרק. החיפוש הנגיש ביותר למאגר הסרטים הוא דווקא דרך הקטלוג – חיפוש התיקים שהמפרסם שלהם הוא "המועצה לביקורת סרטים" ומוביל לרוב התיקים באופן זה.

מתה דאטה:

הפירוט הקיים עבור התיקים מינימלי ואינו מציין את אופי הצנזורה באם הייתה כזו – סך הכול אינו מאפיין את תוכן התיק ודרושה קריאת המסמכים עצמם כדי לדלות מידע. בנוסף, שם הסרט מופיע לפי השם העברי, כלומר התרגום שניתן לו על ידי המועצה, ולכן ייתכנו מצבים רבים שחיפוש לפי שם הסרט ששפת המקור אינה עברית יוביל למבוי סתום. התרגום הוא לאו דווקא מילולי 1:1 אלא לוקח חופש יצירתי ולכן קישור בין הארכיון לבין מקורות חוץ שרובם בשפה האנגלית מהווה אתגר ודורש מעורבות אנוש. קיימים מאגרי סרטים בשפות זרות מפורטים מאוד, וקיימים ממשקים תכנותיים נגישים ופתוחים לגישה ציבורית, מה שמקל על הקישור בכל זאת.

רקע היסטורי:

בישראל קיימת מאז 1927 "המועצה לביקורת סרטים" תחת חקיקת "פקודת סרטי הראינוע" של המנדט הבריטי, ויציבה בתפקידה תחת "תקנות סרטי הראינוע" (84'), "חוק הבזק" (82') וכן "תקנות המועצה לביקורת סרטים ומחזות" (79'), ולפי אתר הממשלה תפקידיה כוללים: מתן אישור לכל סרט המיועד להפצה במרחב הציבורי בכל רחבי הארץ, בין אם בבתי הקולנוע או בכבלים ובטלוויזיה, פיקוח על אותם המוסדות והגבלה על סרט או חלק מסרט שיש בו "חשש נכון להאמין שהוא מוצג שלא ברשות".

בכתבה "המניע להתפטרות" מאת צבי לביא, מתוארת התפטרותו של המשורר חיים גורי בשנת 66' מהמועצה ערב פסיקתה על אישורו של הסרט הישראלי "המניע לרצח", מחששו שתחטא לפסול את הקרנתו. גורי מתאר עצמו "בדרך כלל [...] מאד ליברלי" אך טוען כי "סדיזם וגזענות צריך לאסור". מעניינת אמירתו בכתבה כי "המועצה עומדת לעתים נבוכה בהיעדר קנה מידה לשיפוט". נדמה שגם אותו הטריד החיפוש אחר הדפוס.

שאלת הליברליות של המועצה בזמנים מודרניים מעסיקה ומעוררת ספקות רבות בצורך בה. ב 1982 העלתה "שינוי – מפלגת המרכז" הצעת חוק לביטול הצנזורה על מנת "להסיר מעלינו את הכתם של קיום צנזורה" מעצם טבעה של הדמוקרטיה הישראלית. התומכים גורסים כי עבור אותם מקרים חריגים קיים חוק העונשין בדבר פרסום חומר תועבה, וכי עברה על חוק זה היא פלילית ותשמור על המפתן מפני תכנים שיש קונצנזוס רחב יותר לגבי הצורך במניעתם. בפרוטוקול נראה כי הדיון היה קצר ובסופו תמכו הרוב להעביר לקריאה שנייה, אך כפי שאנו מכירים, הועדה חיה ונושמת.

גם בעידן של זרימת המידע בו אנו חיים הדיון רלוונטי. אכן קל למצוא קישור לצפייה פירטית בכל תוכן שהתפרסם, אך כוחה של הועדה עבר לשלבי ההפקה, למה שנקרא על ידי אנשי הקולנוע הישראלי – "צנזורה עצמית". תקציבי הקולנוע עיקרם בקרנות "מיינסטרימיות" ובתקציבי משרד התרבות. שלבי המיון מתייחסים מטבע הדברים לגורלו של הסרט באולמות הקולנוע – ולכן יש לוודא שיגיע אליהם.

<u>כלים שהיו בשימוש:</u>

סריקת ארכיון המדינה התגלתה כמשימה מרובת רבדים. ראשית היה עלינו ללמוד את המבנה בו האתר מציג תוצאת חיפוש, בפרט כזו של אלפי תיקים. על מנת לדלות את המטה-דאטה של כל תוצאה בנפרד הבנו שיש להשתמש בסלניום, שכן חלק מהאלמנטים הופיעו כתוצאה של הרצת קוד JavaScript. השתמשנו ב - Beautifulsoup לפרסר את הפלט כפי שמחולק במטה דאטה של הארכיון. בשלב זה היה צורך "לנקות" את המידע מפרמטרים שאינם נחוצים או מספקים מידע חדש על ידי פילטרים שיצרנו, וכמובן "לחתוך" את שמות הסרטים מתוספות כדי להכינם להצלבה עם השמות באנגלית. לבסוף על ידי Pandas בנינו מכל התיקים קובץ csv של הארכיון, ובנוסף מיזגנו עם csv נוסף שהכיל בין היתר את שמות הסרטים בעברית וגם באנגלית יחד עם פרמטרים מתוך IMDb שמשכנו בעזרת הספריה IMDbPY.

<u>מה למדנו:</u>

יצירת דיגיטציה של תכנים השייכים לעולם מדעי הרוח היא תחום חדש עבורנו ודורשת הבחנה בפרטים גם עבור מאגרי מידע רב, ובפרט באתר הארכיון שמעלה כמויות של תיקים מידי יום, המגיעים ממשרדים וממקורות שונים ובפורמטים שאינם אחידים. היכולת ליצור קלסיפיקציה יעילה בכל אלה כוללת מעורבות אנוש בנקודות מפתח, אך עם זאת מרבית התהליך היה על ידי קוד ובמרבית המקרים נעשה כהלכה.

דוגמה חשובה הייתה סוגיית התרגום של שמות הסרטים לעברית: במסמכים מופיע התרגום שניתן לשם הסרט על ידי הועדה, אך ברוב המקרים לא מופיע שם המקור. הקישור למקורות חיצוניים דרש את שם הסרט בשפת המקור. לעתים התרגום נעשה באופן יצירתי ואינו בהכרח תרגום סמנטי לכן חיפוש לפי השם בעברית לא בהכרח מוביל לשם הסרט באנגלית. מצאנו כי קל היה יותר להגיע מאוסף של שמות באנגלית שלהם קיים תרגום ומשם ליצור הצלבה עם השמות שנמצאו בתיקי הארכיון, עם זאת ניתן לשים לב כי ישנם כפילויות לסרטים מסוימים. על מנת להסיר כפילויות יש לבדוק את ההתאמה בין הסרט המופיע באתר הארכיון ובין שלל האפשרויות שב IMDb או להפך וזאת עשינו בצורה ידנית שכן אין מספיק פרטים כדי להצליב באופן אוטומטי בין אלה.

<u>מסקנות:</u>

הצורך למצוא פתרונות אוטומטיים לסיווג מידע רב הוא קריטי ועם זאת על מנת לעשות זאת בצורה חכמה - ניכר כי יש מקום למעורבות אנוש, ככל שהקריטריונים לסיווג ברורים יותר מהשלב הראשוני, כך קל יותר בהמשך למצוא נקודות משותפות. אתר הארכיון שמציג מידע רב שנצבר לאורך כל שנות חייה של המדינה הצליח לייצר תבנית בסיסית בצורת המטה דאטה, המהווה נקודת פתיחה טובה למיונים וסיווגים משניים על פי נושאים כפי שהצגנו בעבודתנו בנושא הסרטים. ככל שהחיפוש יהיה תחת קטגוריות ספציפיות ניתן יהיה למצוא מכנים משותפים רבים יותר ולהרחיב את המטה דאטה הקיים. בכל שלב ניסינו למצוא פתרונות אוטומטיים ככל הניתן גם עבור מקרים פרטיים כפי שהצגנו במהלך העבודה, אך היה צורך לבצע סינון נוסף באופן ידני כדי לשמור על מהימנות התוצאות לכן לטעמינו עדיין יש מקום לחיפוש יצירתי אחר שיטות לדיגיטציה של מדעי הרוח.